Report on Multinomial regression using Ridge and Linear Regression for predicting house price

Rahul Mangalampalli

Machine Learning Intern

www.ai-techsystems.com

rahul_mangalampalli@yahoo.in

***Abstract*—In performing data mining, a common task is to search for the most appropriate algorithm(s) to retrieve important information from data. With an increasing number of available data mining techniques, it may be impractical to experiment with many techniques on a specific dataset of interest to find the best algorithm(s). In this paper, we demonstrate the suitability of tree-based multi-variable linear regression in predicting algorithm performance. We take into account prior machine learning experience to construct metaknowledge for supervised learning. The idea is to use summary knowledge about datasets along with past performance of algorithms on these datasets to build this meta-knowledge. We augment pure statistical summaries with descriptive features and a misclassification cost, and discover that transformed datasets obtained by reducing a high dimensional feature space to a smaller dimension still retain significant characteristic knowledge necessary to predict algorithm performance. Our approach works well for both numerical and nominal data obtained from real world environments.**

Keywords*—*;* *regression; dimensionality reduction; combined metric

I. Introduction

Multiple regression is an extension of simple linear regression. It is used when we want to predict the value of a variable based on the value of two or more other variables. The variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable). The variables we are using to predict the value of the dependent variable are called the independent variables (or sometimes, the predictor, explanatory or regressor variables).

For example, you could use multiple regression to understand whether exam performance can be predicted based on revision time, test anxiety, lecture attendance and gender. Alternately, you could use multiple regression to understand whether daily cigarette consumption can be predicted based on smoking duration, age when started smoking, smoker type, income and gender.

Multiple regression also allows you to determine the overall fit (variance explained) of the model and the relative contribution of each of the predictors to the total variance explained. For example, you might want to know how much of the variation in exam performance can be explained by revision time, test anxiety, lecture attendance and gender "as a whole", but also the "relative contribution" of each independent variable in explaining the variance.

This "quick start" guide shows you how to carry out multiple regression using SPSS Statistics, as well as interpret and report the results from this test. However, before we introduce you to this procedure, you need to understand the different assumptions that your data must meet in order for multiple regression to give you a valid result. We discuss these assumptions next.

II. Linear Multiple Regression

The general purpose of multiple regression (the term was first used by Pearson, 1908) is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable. For example, a real estate agent might record for each listing the size of the house (in square feet), the number of bedrooms, the average income in the respective neighborhood according to census data, and a subjective rating of appeal of the house. Once this information has been compiled for various houses it would be interesting to see whether and how these measures relate to the price for which a house is sold. For example, you might learn that the number of bedrooms is a better predictor of the price for which a house sells in a particular neighborhood than how "pretty" the house is (subjective rating). You may also detect "outliers," that is, houses that should really sell for more, given their location and characteristics.

Personnel professionals customarily use multiple regression procedures to determine equitable compensation. You can determine a number of factors or dimensions such as "amount of responsibility" (*Resp*) or "number of people to supervise" (*No_Super*) that you believe to contribute to the value of a job. The personnel analyst then usually conducts a salary survey among comparable companies in the market, recording the salaries and respective characteristics (i.e., values on dimensions) for different positions. This information can be used in a multiple regression analysis to build a regression equation of the form:

Salary = .5*Resp + .8*No_Super

Once this so-called regression line has been determined, the analyst can now easily construct a graph of the expected (predicted) salaries and the actual salaries of job incumbents in his or her company. Thus, the analyst is able to determine which position is underpaid (below the regression line) or overpaid (above the regression line), or paid equitably.

In the social and natural sciences multiple regression procedures are very widely used in research. In general, multiple regression allows the researcher to ask (and hopefully answer) the general question "what is the best predictor of ...". For example, educational researchers might want to learn what are the best predictors of success in high-school. Psychologists may want to determine which personality variable best predicts social adjustment. Sociologists may want to find out which of the multiple social indicators best predict whether or not a new immigrant group will adapt and be absorbed into society.

III. Ridge Regression

Ridge regression is a way to create a parsimonious model when the number of predictor variables in a set exceeds the number of observations, or when a data set has multicollinearity (correlations between predictor variables).

Tikhivov’s method is basically the same as ridge regression, except that Tikhonov’s has a larger set. It can produce solutions even when your data set contains a lot of statistical noise (unexplained variation in a sample).

A.Ridge Regression vs. Least Squares

Least squares regression isn’t defined at all when the number of predictors exceeds the number of observations; It doesn’t differentiate “important” from “less-important” predictors in a model, so it includes all of them. This leads to overfitting a model and failure to find unique solutions. Least squares also has issues dealing with multicollinearity in data. Ridge regression avoids all of these problems. It works in part because it doesn’t require unbiased estimators; While least squares produces unbiased estimates, variances can be so large that they may be wholly inaccurate. Ridge regression adds just enough bias to make the estimates reasonably reliable approximations to true population values.

B.Shrinkage

Ridge regression uses a type of shrinkage estimator called a ridge estimator. Shrinkage estimators theoretically produce new estimators that are shrunk closer to the “true” population parameters. The ridge estimator is especially good at improving the least-squares estimate when multicollinearity is present.

C.Regularization

Ridge regression belongs a class of regression tools that use L2 regularization. The other type of regularization, L1 regularization, limits the size of the coefficients by adding an L1 penalty equal to the absolute value of the magnitude of coefficients. This sometimes results in the elimination of some coefficients altogether, which can yield sparse models. L2 regularization adds an L2 penalty, which equals the square of the magnitude of coefficients. All coefficients are shrunk by the same factor (so none are eliminated). Unlike L1 regularization, L2 will not result in sparse models.

A tuning parameter (λ) controls the strength of the penalty term. When λ = 0, ridge regression equals least squares regression. If λ = ∞, all coefficients are shrunk to zero. The ideal penalty is therefore somewhere in between 0 and ∞.

D.On Mathematics

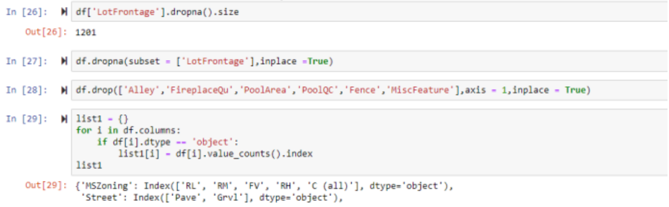
OLS regression uses the following formula to estimate coefficients:
ridge regression

If X is a centered and scaled matrix, the crossproduct matrix (X`X) is nearly singular when the X-columns are highly correlated. Ridge regression adds a ridge parameter (k), of the identity matrix to the cross product matrix, forming a new matrix (X`X + kI). It’s called ridge regression because the diagonal of ones in the correlation matrix can be described as a ridge. The new formula is used to find the coefficients:

Choosing a value for k is not a simple task, which is perhaps one major reason why ridge regression isn’t used as much as least squares or logistic regression. You can read one way to find k in Dorugade and D. N. Kashid’s paper Alternative Method for Choosing Ridge Parameter for Regression..

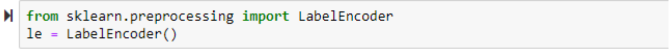
IV. Impelementing using Python

First of all,data has been cleaned using various data cleaning processes like dropping n/a rows and columns.
Secondly columns which have very less no of n/a values have replaced using mean and median values.

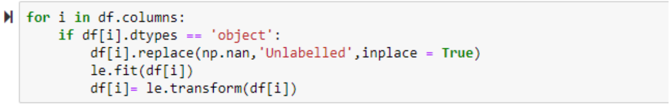


```
In [26]:  df['LotFrontage'].dropna().size
Out[26]: 1201

In [27]:  df.dropna(subset = ['LotFrontage'],inplace =True)

In [28]:  df.drop(['Alley','FireplaceQu','PoolArea','PoolQC','Fence','MiscFeature'],axis = 1,inplace = True)

In [29]:  list1 = {}
          for i in df.columns:
              if df[i].dtype == 'object':
                  list1[i] = df[i].value_counts().index
          list1
Out[29]: {'MSZoning': Index(['RL', 'RM', 'FV', 'RH', 'C (all)'], dtype='object'),
          'Street': Index(['Pave', 'Grvl'], dtype='object'),
```

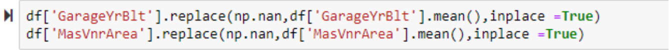
Then columns which can be encoded in categorical values have been changed for easy prediction.



```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
```



```
for i in df.columns:
    if df[i].dtypes == 'object':
        df[i].replace(np.nan,'Unlabelled',inplace = True)
        le.fit(df[i])
        df[i]= le.transform(df[i])
```



```
df['GarageYrBlt'].replace(np.nan,df['GarageYrBlt'].mean(),inplace =True)
df['MasVnrArea'].replace(np.nan,df['MasVnrArea'].mean(),inplace =True)
```

Then the whole data has been used to predict values of house prices which showed very less difference between both the algorithms.
Link of my repository:https://github.com/rahulmangalampalli?tab=repositories